Summary of the Fall 2006 HEPiX Meeting

Technology Meeting
October 23, 2006

Alexander Withers, alexw@bnl.gov

BROOKHAVEN NATIONAL LABORATORY

Site Reports Talks

Site Reports

- CERN
  - SLC4 for LHC startup (SLC5 probably too late)
  - Experienced power and A/C failures
  - 30 dual-core AMD, 100 dual-core Woodcrests
  - Using castor2 for distributed storage
  - Oracle RAC for physics DB
  - New tape robots from Sun and IBM tested

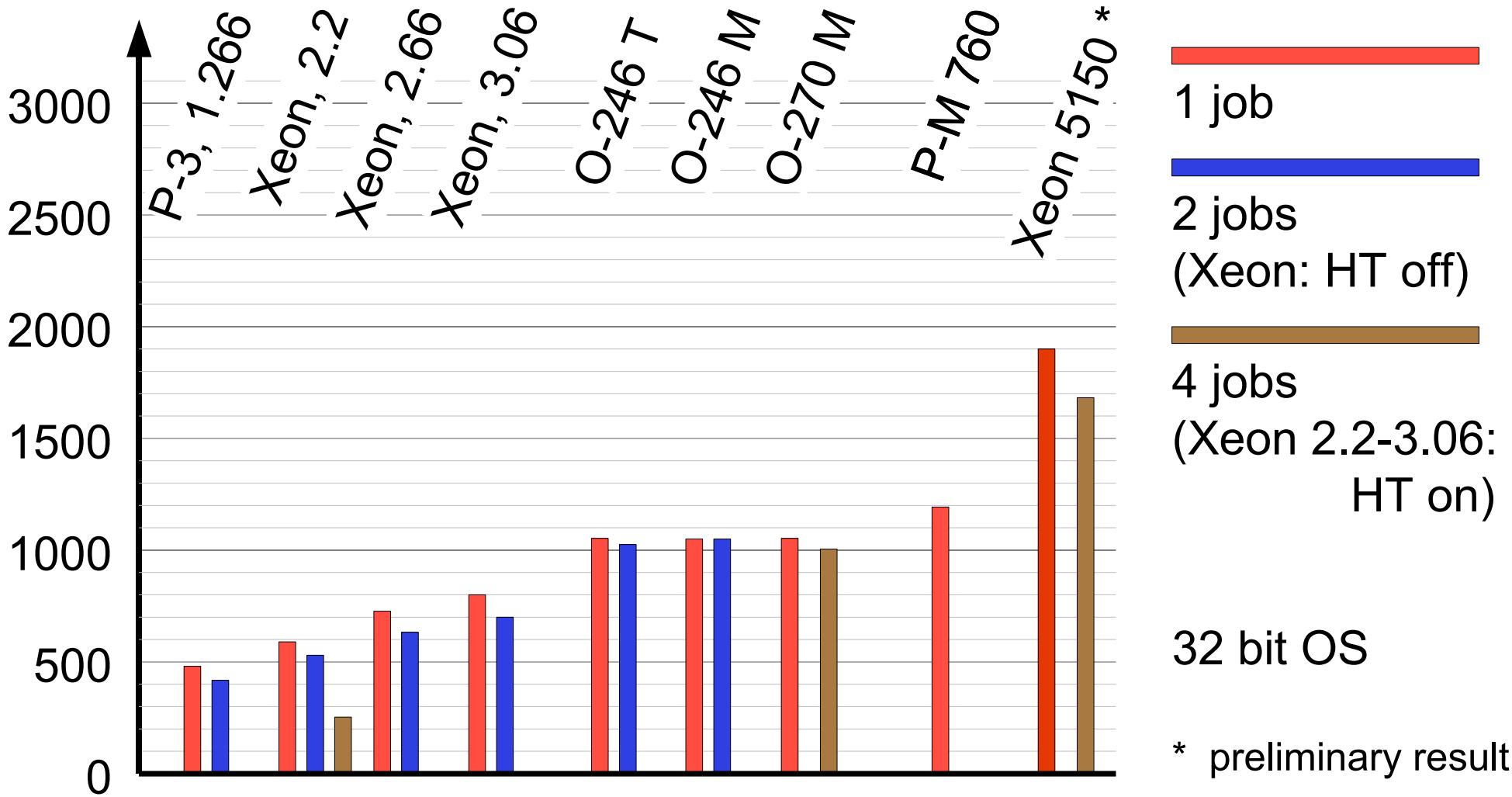
Site Reports

- Fermilab
  - Experiencing same security situation as BNL
  - 2000 sq. ft. facility for 3080 machines, another one being built
  - One STK SLA8500, another on order
    - 4PB on tape, expect 4-8PB per year, peaks of >25TB per day to&from tape
  - Farms
    - Dedicated clusters for CDF, D0, USCMS, LQCD, and SDSS
    - General purpose farm is entry point for FermiGrid/OSG.
  - USCMS
    - 700TB dCache, 2.5PB by fall 2007
    - Bluearc for online filesystems (two heads)
    - 700 node farm, 1600 by fall 2007

Site Reports

- GridKa
  - 3 year contract with NEC for 4500TB, already have 500
  - dCache pool on GPFS
  - 153 dual-core opteron 270
  - Problems with water cooling system leaking
  - Benchmarks
    - SPECint2000 with gcc -O3 -march -funroll-loops
    - “one benchmark per CPU core in parallel”
    - Intel Xeon 5150 Woodcrest 2.66 GHz

CPU Speed (SPECint_base2000 *per Job*)



Site Reports

- JLab
  - New IT department with Roy Whitney at the head
  - RHEL4, moving to 64bit
  - Glue-X collaboration to use OSG as a baseline
  - Integration of clusters (LSF and PBS)
  - 10 GigE networking upgrade
  - Same DOE security issues as BNL and Fermi
    - Two-factor authentication

Site Reports

- NERSC
  - Security: one time passwords
  - Working on a HSM for GPFS
  - Using nagios for their farm, examining cacti and ganglia
  - 19,000+ processor Cray cluster
  - Power: 2 megawatts, 5.5 expected
  - Cooling: air cooled, 25 megawatts of cooling expected

Site Reports

- TRIUMF
  - Networking infrastructure to teir-2 sites being expanded
  - Expect 10Gb lightpath to CERN Nov 2006
  - "EtherDrive" system: SATA storage over ethernet
    - Seen by Linux as a block device
  - Using dCache
  - Moving to SL4
  - Blades favored over 1U: power and space savings
  - A/C failures

Site Reports

- NIKHEF
  - Storage
    - 4TB of EMC storage, using snapshots for home directories
    - iSCSI being tested
    - NFS works well with SLC4 but not SLC3
  - Using SLC3/4
  - Need 300K SPECint2000s w/20GB RAM per core, currently has 34 Dell 1950s

Site Reports

- RAL
  - 52 dual-core AMD opteron 270 w/4GB RAM
  - 168TB SATA storage, 282TB being procured
  - 64 dual-core Woodcrests being procured w/4GB RAM
  - A/C failures --> machine failures --> alarm system failure
  - Moving to SL4 from SL3 and RH7.2
  - Using dCache but moving to castor2
  - Ganglia and Nagios for monitoring

Site Reports

- INFN
  - GPFS being used (trouble with SL3 as NFS server)
  - Moving to LSF for batch computing
  - Upgrades of power and cooling
  - Testing iSCSI as a FC replacement
    - Only getting < 70MB/s
  - Moving to LDAP authentication/authorization

Site Reports

- GSI
  - Mass storage
    - "d" filesystem for GSI
      - Based on xrootd and GSTORE
      - In testing phase
  - Debian on 800 CPUs
  - Testing quad dual-core systems w/32Gb RAM on Debian

Site Reports

- DAPNIA
  - 23 new IBM worker nodes
  - 16x400GB of SCSI storage
  - Using Quattor and Lustre
- LAL
  - 10TB of storage on Tru64
  - 30 cores for computing, getting 25 dual-core Woodcrests
    - Cheaper than opteron
  - Moving to SL4.3 from SL3, 66% done
  - Using Quattor, testing xrootd

Site Reports

- SLAC
  - Now a ATLAS teir-2 site
  - New 153 node cluster: dual-core opteron 2.3 GHz
  - 755TB of Sun-based SATA storage
  - Testing Lustre: goal of 1GB/s over Infiniband for chkpointing
  - Using RHEL3/4 and SL3, moving to SL4 due to ATLAS
  - Using both 32 and 64 bit architectures
  - Use of .k5login to supersede ssh keys
  - Working on serial console server with CERN\
    - GSSAPI, ACLs, and heartbeats

Site Reports

- INFN-CNAF
  - Cooling problems, A/C failure
  - Farm
    - 1.5M SI2K
    - Quattor and planning on Lemon
    - Upgrading SLC4 from SLC3
      - Need SLC4 for Woodcrest?
    - LSF
  - Storage: castor, GPFS, StoRM (SRM layer for GPFS), 600TB total, 400TB more by end of 2006

Non-Site Report Talks

Infrastructure

- SL update
  - 18000+ computers with SL (from ftp logs)
  - SL4.4 released in October
  - XFS not in the main release, still in contrib
    - 32bit kernel module still buggy
  - SL5
    - Working on installer changes
    - Removing trademarks
    - Not based on RHEL5 Beta1
  - Planning on SL4.5 when RHEL4.5 comes out

Infrastructure

- CERN's talk on Service Level Status
  - Monitoring app that works with other monitoring apps
  - Not a monitoring framework
  - Measures quality of service not just availability

Infrastructure

- SLAC's talk on workflow management with RT
  - Workflow problems: completed tasks would go unnoticed, poor communication between groups
  - Already using RT which has requests, owners, dependencies, and time tracking
  - One queue for installation requests
    - Each ticket has various subtasks (child tickets)
    - Child tickets are assigned to different groups
    - Updated child ticket will update the parent ticket
  - Model gives a concise overview of what's waiting for what
  - Using the same model for a planning queue

Infrastructure

- GSI's talk on high availability
  - Examined commercial methods from Microsoft and Oracle
  - Went with OSS solution
    - **Heartbeat** to trigger fail over when service is down
    - **DRBD** for raid1 over network
      - Kernel module, sits between filesystem and disk
    - **MON** for monitoring

Security

- FNAL's talk on SL inventory project
  - Software inventory solution
  - Configuration management
  - Provides hardware and software management for all computers
  - Audits
    - Packages installed
    - Non-standard packages
    - Hardware installed
    - Network configuration, network state, etc.

Mass Storage

- JLab's talk on mass storage
  - Jasmine manages the mass storage
  - StorageTek drives attached to Linux movers
  - Data is always buffered
  - Read in order and write in batches to minimize tape mounts
  - Multi-level disk cache
    - L1 must match speed of tape
    - L2: multiple streams aggregated to L1
    - Client aggregated to L2

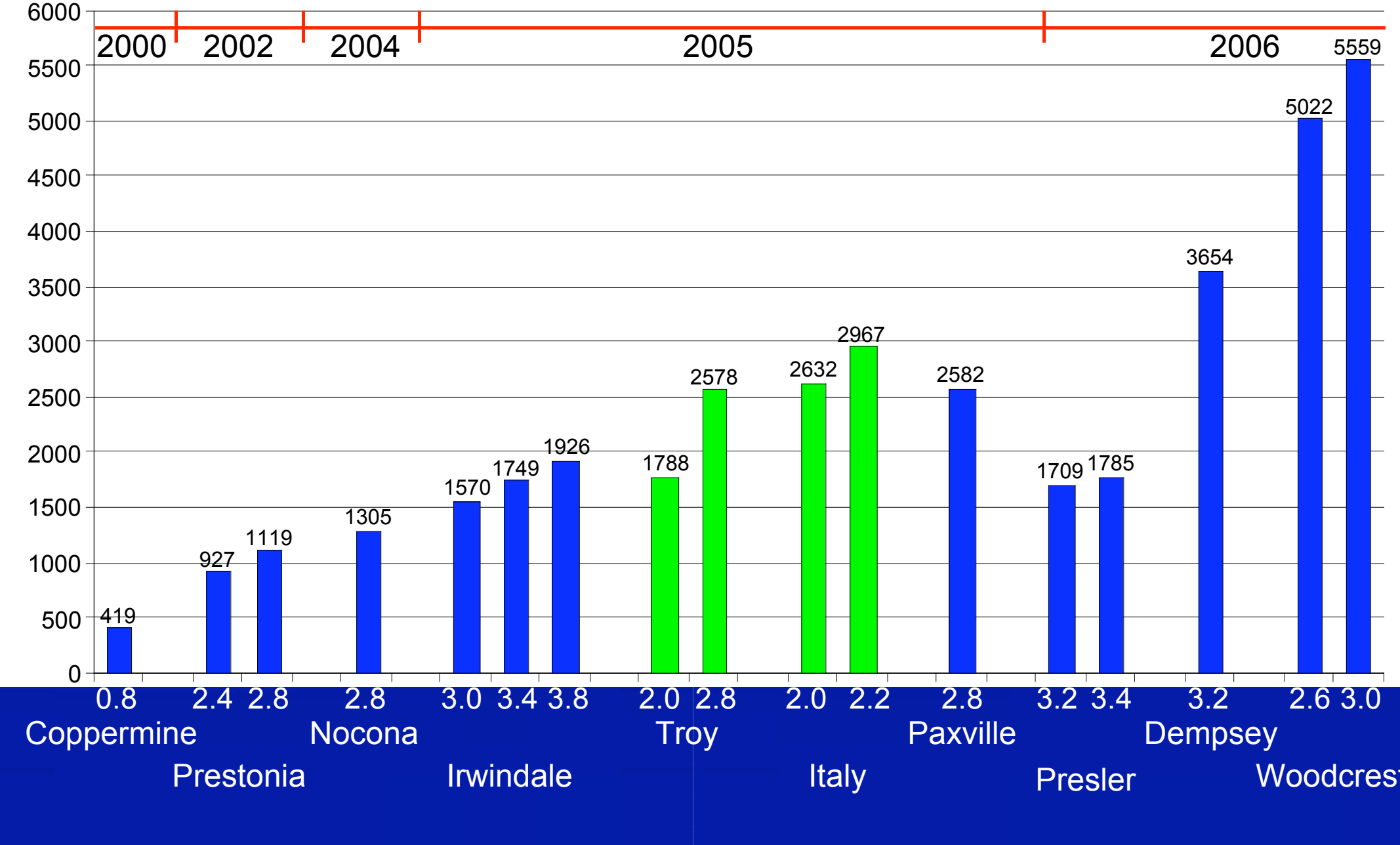
Mass Storage

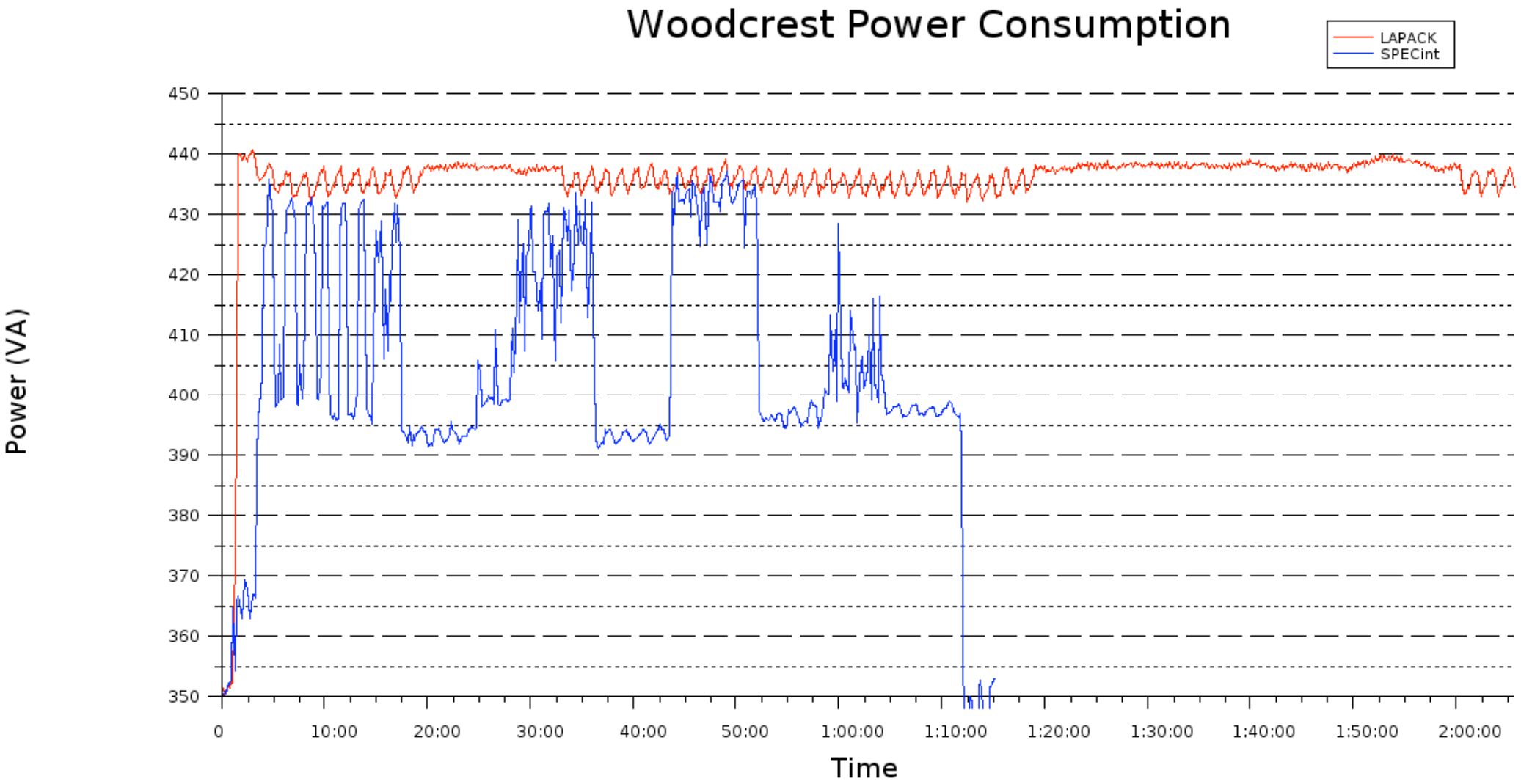
- CCIN2P3's talk on storage classes
  - Improve usage quality of storage resources
  - Three classes
    - On tape, not on disk
    - Only on disk
    - On disk and on tape
  - One storage class per namespace

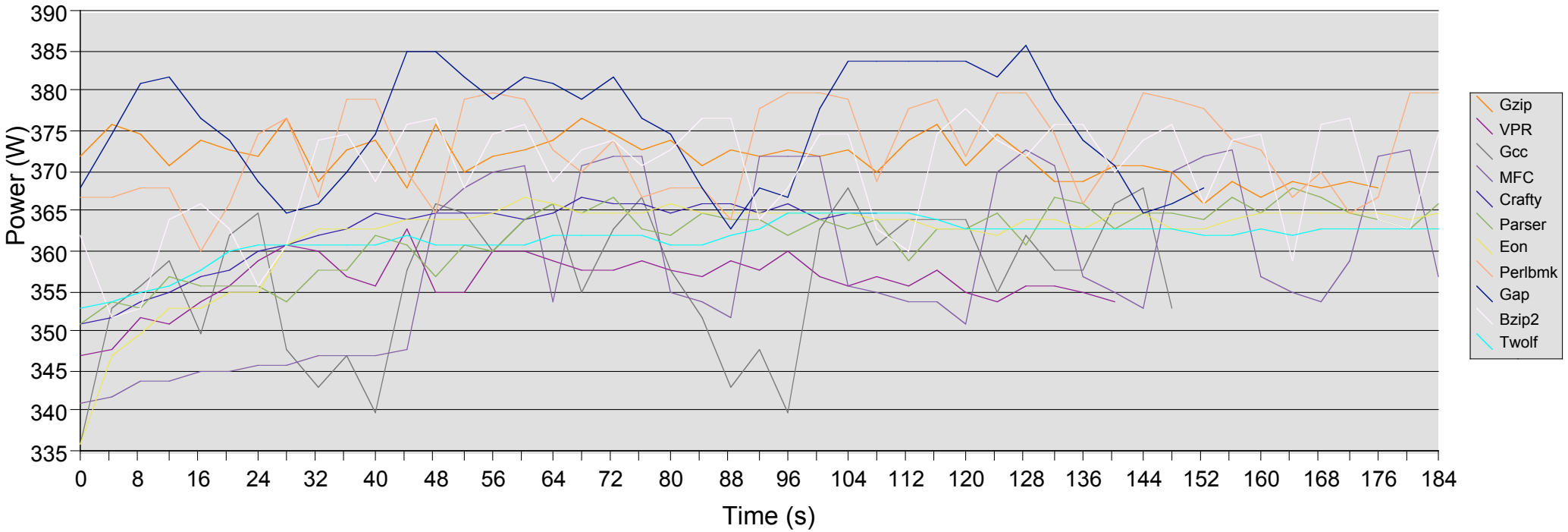
Benchmarking

- CERN's talk on benchmarking
  - SPECint fits CERN's usage pattern, looking at ways to quantify this
  - Want the benchmark to give a realistic value of the machine
  - Tells vendors they want 500K SPECints, not 300 Xeon CPUs
    - Gives vendors the scripts and benchmarking software
    - Works well
  - Power consumption
    - Now being considered
    - Measures the primary AC power circuit: W and VA
    - Creates realistic conditions with LAPACK (SPECint does not provide a stable load)

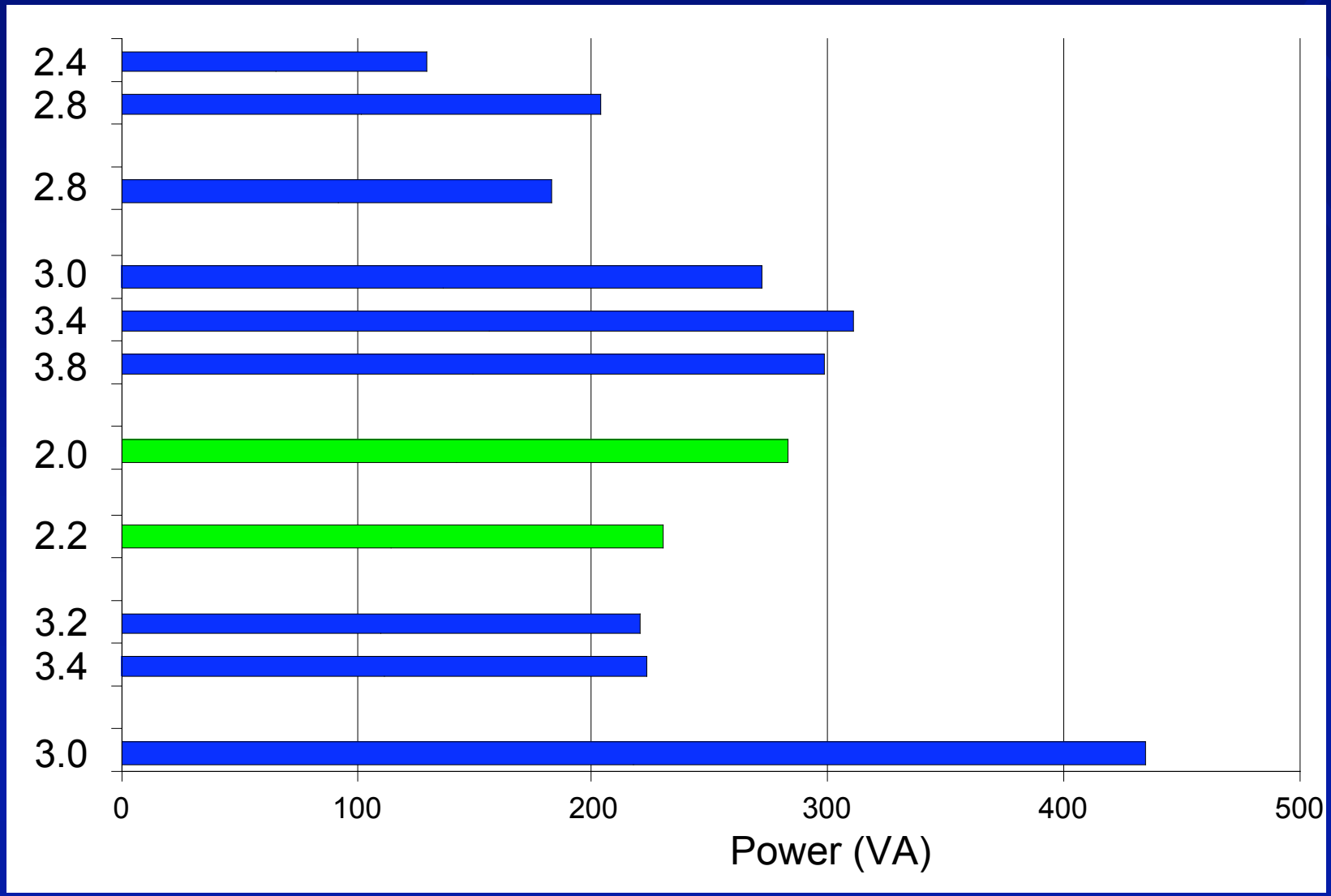
Evolution of SPECint at CERN




Woodcrest Power Consumption
LAPACK
SPECint
Power (VA)
450
440
430
420
410
400
390
380
370
360
350
0
10:00
20:00
30:00
40:00
50:00
1:00:00
1:10:00
1:20:00
1:30:00
1:40:00
1:50:00
2:00:00
Time

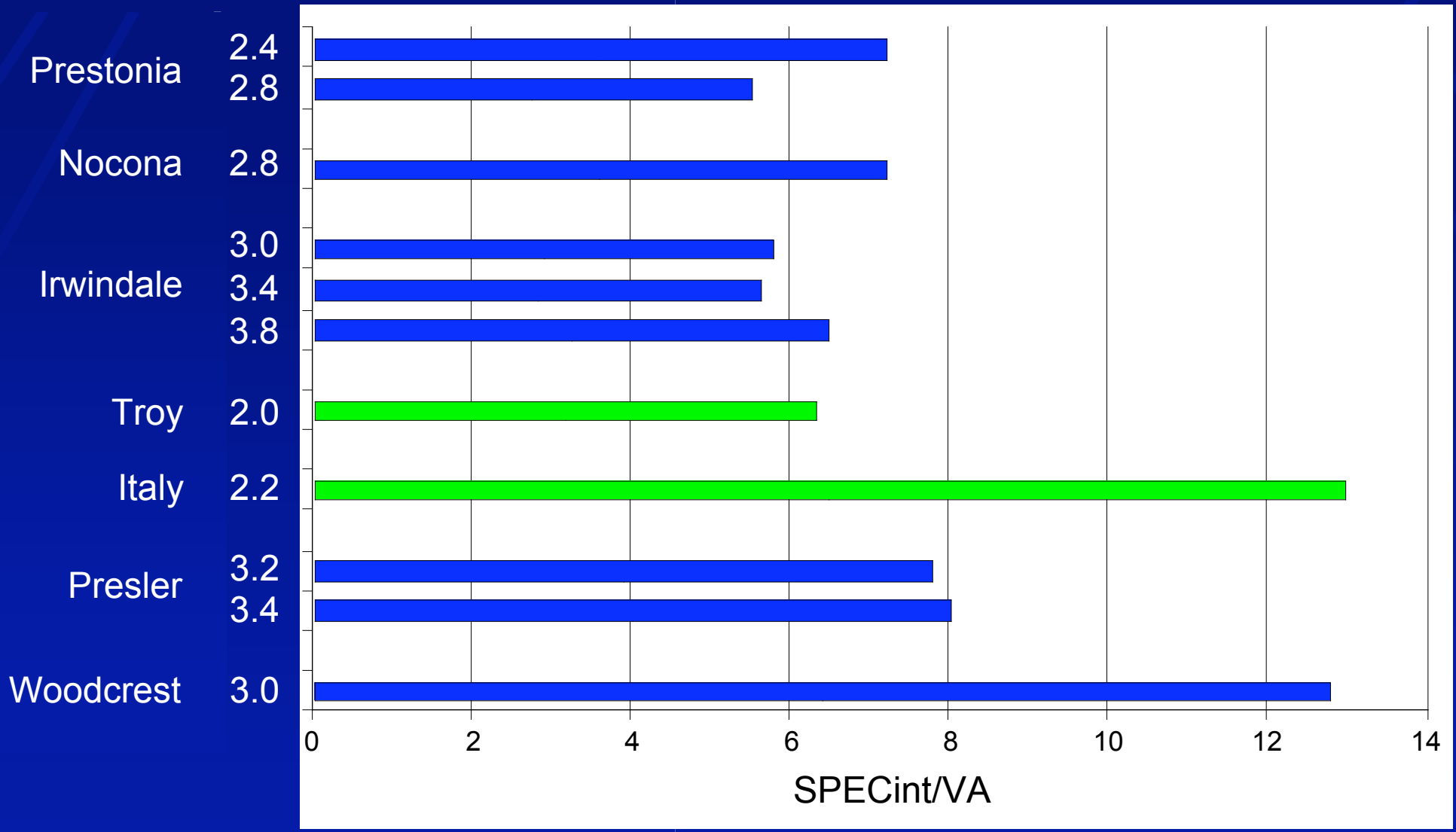

Power (W)
390
385
380
375
370
365
360
355
350
345
340
335
0
8
16
24
32
40
48
56
64
72
80
88
96
104
112
120
128
136
144
152
160
168
176
184
Time (s)
Gzip
VPR
Gcc
MFC
Crafty
Parser
Eon
Perlbmk
Gap
Bzip2
Twolf

Power Consumption: Results



Systematic error ≈ 5 - 10%

SPECint per VA



Infrastructure

- CERN's fabric management talk
  - LSF and Grid integration
    - Two methods for CPU queue selection: LCG and gLite
    - One method for Memory queue selection: gLite
  - SMART monitoring
    - Includes metrics and self-tests
    - OSS tools available to access SMART data
    - Proactively replacing disks, some vendors willing

Grid Technologies

- Talk from ACIS/UofF on virtual machines
  - Using virtual machines for providing proper execution environment
  - Middleware takes care of instantiating VM
  - Also: virtual networks and virtual storage
  - In-VIGO: software to integrate VM technology into the grid
    - Hides complexity from users
    - Provide secure execution environments

Grid Technologies

- OSG progress and vision by Keith Chadwick
  - >15000 CPUs, 6PB MSS, 4PB disk, 27 VOs
  - Co-funded by DOE and NSF for 5 years @ $6M/year
  - Support for non-Physics communities
  - Improvements to VDT
    - Add dCache/SRM
  - OSG 0.6.0 in early 2007, relies on VDT 1.4.0

OSG Middleware

Open Science Grid

- OSG Middleware is deployed on existing farms and storage systems.
- OSG Middleware interfaces to the existing installations of OS, utilities and batch systems.
- VOs have VO scoped environments in which they deploy applications (and other files), execute code and store data.
- VOs are responsible for and have control over their end-to-end distributed system using the OSG infrastructure.

Applications

User Science Codes and Interfaces

VO Middleware

ATLAS Panda, DQ etc

Bio blast, charmm etc.

LIGO LDR, Catalogs etc.

CMS cmssw, LFC, etc.

Infrastructure

OSG Release Cache:

VDT + OSG specific configuration + utilities.

Virtual Data Toolkit (VDT)

core technologies + software needed by stakeholders:
e.g.VOMS, CEMon VDS, MonaLisa, VO-Privilege.

Core grid technology distributions:

Condor, Globus, Myproxy

Existing Operating, Batch systems and Utilities.

Grid Technologies

- Other Grid talks:
  - GridX1: Canadian grid for HEP applications (I. Gable)
  - GridPP (John Gordon)
  - EGEE Grid Infrastructure (Ian Bird)
  - Issues and problems around Grid site management (Ian Bird)
  - Grid Security in WLCG and EGEE (David Kelsey)

Other Talks of Interest

- Experiences with SpamCop from Fermilab
- TRAC: wiki-based issue tracker from LAL/IN2P3
- TWiki talk by CERN
- Using Quattor to manage grid infrastructure by LAL/IN2P3
- GSI's experiences fighting spam
- KRB5 and Torque by Desy
- Nice talk on 64bit computing by JLab
- Network security monitoring with Squil by JLab
- Stakkato intrusions